Plant Signaling & Behavior 7:7, 864–873; July 2012; © 2012 Landes Bioscience

# Bimodal expression level polymorphisms in *Arabidopsis thaliana*

Atsushi J. Nagano,[1,*] Takashi Tsuchimatsu,[2,3] Yudai Okuyama[4] and Ikuko Hara-Nishimura[5,*]

[1]Center for Ecological Research; Kyoto University; Otu, Shiga, Japan; [2]Department of General Systems Studies; Graduate School of Arts and Sciences; University of Tokyo; Komaba, Tokyo, Japan; [3]Department of Evolutionary Functional Genomics; Institute of Plant Biology; University of Zurich; Zurich, Switzerland; [4]Tsukuba Botanical Garden; National Museum of Nature and Science; Ibaraki, Japan; [5]Department of Botany; Graduate School of Science; Kyoto University; Kyoto Japan

**Keywords:** expression level polymorphism, ELP, transcriptome, bELP, nucleotide diversity, linkage disequilibrium

Differences in gene expression are termed expression level polymorphisms (ELPs). Here, we propose a new ELP class, bimodal ELPs (bELPs), as a criterion to screen for genes that are responsible for natural phenotypic variation and/or that are targeted by balancing selection. bELP genes are characterized by two expression level modes. Genomic scans based on nucleotide sequences are not ideal for identifying genes targeted for selection. A critical concern is that several genes can be present in the selection-targeted regions identified by such scans. This situation indicates the importance of integrating genomic sequence data and other information, such as gene expression data. Comparative transcriptomics is useful for determining evolutionarily and ecologically important polymorphisms. In a genome-wide expression screen of 34 accessions, we identified 344 *Arabidopsis thaliana* genes exhibiting bELPs. Population genetic analysis revealed that bELP genes had high nucleotide diversities and long linkage disequilibriums. The highest nucleotide diversity (11-fold greater than the genomic mean) was found in the At1g23780 gene, which encodes a putative F-box protein. We observed a clear association between the expression mode and sequence type of the At1g23780 gene. Our results suggest that bELPs will be useful for the screening and functional analysis of genes responsible for phenotypic polymorphisms. Such a "multi-omics" approach has the potential to facilitate the scanning of genes relevant to balanced polymorphisms not only in *A. thaliana*, but also in other model and non-model organisms.

## Introduction

Balanced polymorphisms are characterized by two or more alleles that are selectively maintained within populations or species.[1-3] Several selective mechanisms that can maintain polymorphisms have been proposed, such as frequency-dependent selection, spatially and temporally unstable selection and heterozygote advantage.[2] Balanced polymorphisms are associated with specific levels and patterns of nucleotide variation at the selective target and linked genomic regions. Thus, the molecular signature of a given locus can be used to detect an adaptation. This signature can include intermediate frequency polymorphisms and a peak of increased nucleotide diversity.[4] Molecular evidence for balanced polymorphisms has been obtained for several genes and/or genomic regions, e.g., the human class I and II *MHC* genes,[5] Drosophila *Adh*,[6,7] *Arabidopsis thaliana MAM*[8] and disease resistance loci.[9-12]

Genomic scans based on nucleotide sequences are used to identify genes targeted by selection. Even though nucleotide polymorphisms are the primary genetic cause of phenotypic variations, these genomic scans have a drawback in that numerous genes can be present in selection-targeted regions identified by genomic scans. A gene targeted by selection cannot be identified from other candidate genes except in cases where a clear peak of nucleotide diversity is observed, because linkage disequilibrium (LD) and/or a stretch of high nucleotide diversity spans a large genomic region that contains many candidate genes.[13,14] In such cases, additional information, such as expression data, is needed to determine the true selection target. Expression data provides clues about functional polymorphisms as expression level polymorphisms (ELPs). ELPs are defined as differences in gene expression between individuals.[15,16] The levels of ELPs in a genome macroscopically correlate with the levels of nucleotide diversity.[15] ELPs have been shown to contribute to inter- and intra- specific phenotypic variation.[17] Examples of such variation in plants include the liberation of the kernel from the hardened casing, as observed in maize domestication,[18] the loss of seed shattering in rice domestication,[19] flowering time control,[20-22] pathogen resistance,[23,24] insect resistance and secondary metabolism in Arabidopsis.[25-27] ELPs can be caused by various types of DNA sequence polymorphisms, including those affecting trans-acting factors, cis-acting promoter regions, cis-acting splice sites and whole or partial gene deletions.

Recent advances in molecular genetic techniques make it possible to identify loci responsible for adaptation and to study how these loci are selected.[28-30] In particular, identifying genes targeted by balancing selection and unveiling the genetic

*Correspondence to: Atsushi J. Nagano and Ikuko Hara-Nishimura; Email: anagano@gr.bot.kyoto-u.ac.jp and ihnishi@gr.bot.kyoto-u.ac.jp
Submitted: 04/15/12; Accepted: 04/23/12
http://dx.doi.org/10.4161/psb.20486

mechanisms that give rise to natural variation are important for both basic science (e.g., the functional analysis of biological mechanisms) and applied science (e.g., molecular breeding assisted by genomics).[31] In most previous studies, the genes responsible for natural phenotypic variation and/or the genes that are targeted by balancing selection have been identified by a combination of genetic mapping (e.g., QTL mapping and positional cloning) and functional molecular analysis of candidate genes (e.g., expression analysis and transgenic assay).[32,33] This situation indicates the importance of integrating multiple data from various experiments.

In this article, we propose a new class of ELPs, bimodal expression level polymorphisms (bELPs), as a criterion for screening for candidate genes targeted by balancing selection. bELP genes are characterized by two distinct expression levels. In other words, two modes are distinguishable in a histogram of expression levels of bELP genes (**Fig. 1A**). A bELP is an associated concept with a large effect eQTL (expression QTL); some of the bELPs found in natural populations may also be identified by quantitative genetics using mapping populations, and vice versa.[34-36] We have identified genes that exhibit bELPs in the *A. thaliana* genome. Sequencing and population genetic analysis of bELP genes revealed that they have significantly greater nucleotide diversity and longer blocks of linkage disequilibrium than the genomic average. Thus, regions containing bELP genes exhibit characteristics of balancing selection. Our results suggest that a "multi-omics" approach, namely the integration of bELP screening with evolutionary population genomics, will provide a good starting point to conduct a further functional analysis of balancing selection.

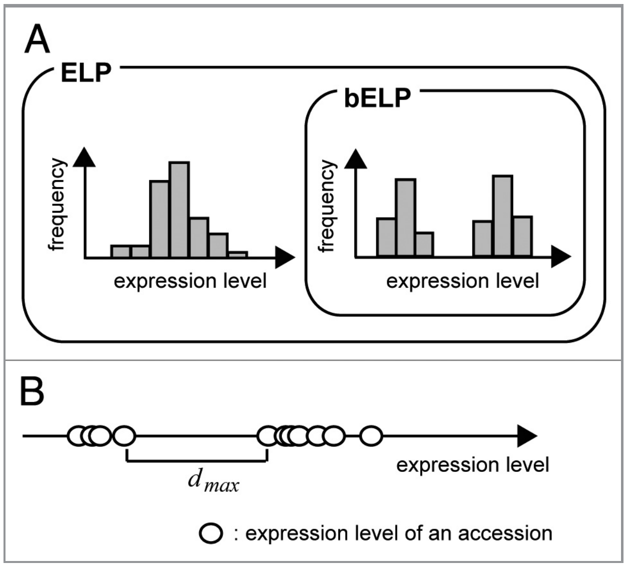

**Figure 1.** Schematic representation of bELPs. (A) Histograms depicting typical ELPs (left) and bELPs (right). In the bELP histogram, two expression modes are clearly distinguishable. The ELP gene superclass includes the class of bELP genes. (B) Screening criterion for identifying bELP genes. Expression levels of a specific gene in each accession are plotted (circles) on an axis. $d_{max}$ is the greatest difference observed between the expression levels of two neighboring accessions.

## Results

**Screening for bELPs in the *A. thaliana* genome.** The A. thaliana genome was first screened for genes exhibiting bELPs using a publicly available DNA microarray data set (see Materials and Methods). At first, we tried to fit the expression data of each gene to a mixture distribution of two normal distributions (**Fig. S1**). However, we realized that it was difficult to obtain reasonable results using this approach, because of insufficient data to estimate the distribution. Thus, we decided to employ a simple non-parametric strategy for the screening (see Materials and Methods). Using lower thresholds, more genes were identified as bELP genes (103 to 872 genes, **Table 1**). The ratio of singletons (71% to 73%) did not depend on the threshold. We used $d_{max}$ = 1 as the threshold for the following analysis, because it corresponds to a 2-fold change, which is often designated as a threshold in empirical studies using DNA microarrays. A 2-fold difference between levels of expression was expected to be clearly distinguishable when using the DNA microarray and real-time PCR techniques. With this threshold, 344 genes were identified as bELP genes (**Fig. 2A**; **Fig. S2 and Table S3**). The frequency of two expression types varied from 2:32 to 17:17 (**Fig. 2B**). To assess whether sequence polymorphisms on the probe positions affect the apparent expression level, we sequenced one of the bELP genes, At1g23780, in 22 accessions (**Fig. S3**). We identified 13 SNPs on 7 of 11 probe positions. The exact same set of SNPs was found in both the high and low expression type. This showed that the SNPs on the probe positions did not affect the apparent expression level of the At1g23780 gene. In addition to this result, it was estimated that SFPs (single feature polymorphisms) were responsible for ~0.13% of the apparent variance in expression level in the previous study using Col-0 and Cvi-1.[15] Therefore, we concluded that most of the bELPs were true ELPs, and not the result of mis-hybridization by sequence polymorphisms. The bELP genes included a large number of biotic stress-related genes (GO: 0030383 host-pathogen interaction, $p = 3.4 \times 10^{-5}$; and GO: 0003793 defense immunity protein activity, $p = 1.0 \times 10^{-4}$) and genes whose products are located in endomembrane systems (GO: 0012505 endomembrane system, $p = 2.8 \times 10^{-5}$). The bELP genes included some genes known to be targets of balancing selection, e.g., ESP and At1g63880 (**Fig. 2C**), as detailed in the Discussion.

**Population genetic analysis of 20 bELP genes.** To determine the level of nucleotide diversity (π) of bELP genes, genomic sequences of 20 bELP genes were obtained from eight to 23

**Table 1.** Results of bELP screening with various thresholds

| Threshold | | bELP | | |
|---|---|---|---|---|
| Log2 | Fold change | With singleton | Without singleton | Ratio of singleton (%) |
| 0.58 | 1.5 | 2982 | 872 | 71 |
| 1 | 2 | 1276 | 344 | 73 |
| 1.58 | 3 | 580 | 165 | 72 |
| 2 | 4 | 370 | 103 | 72 |

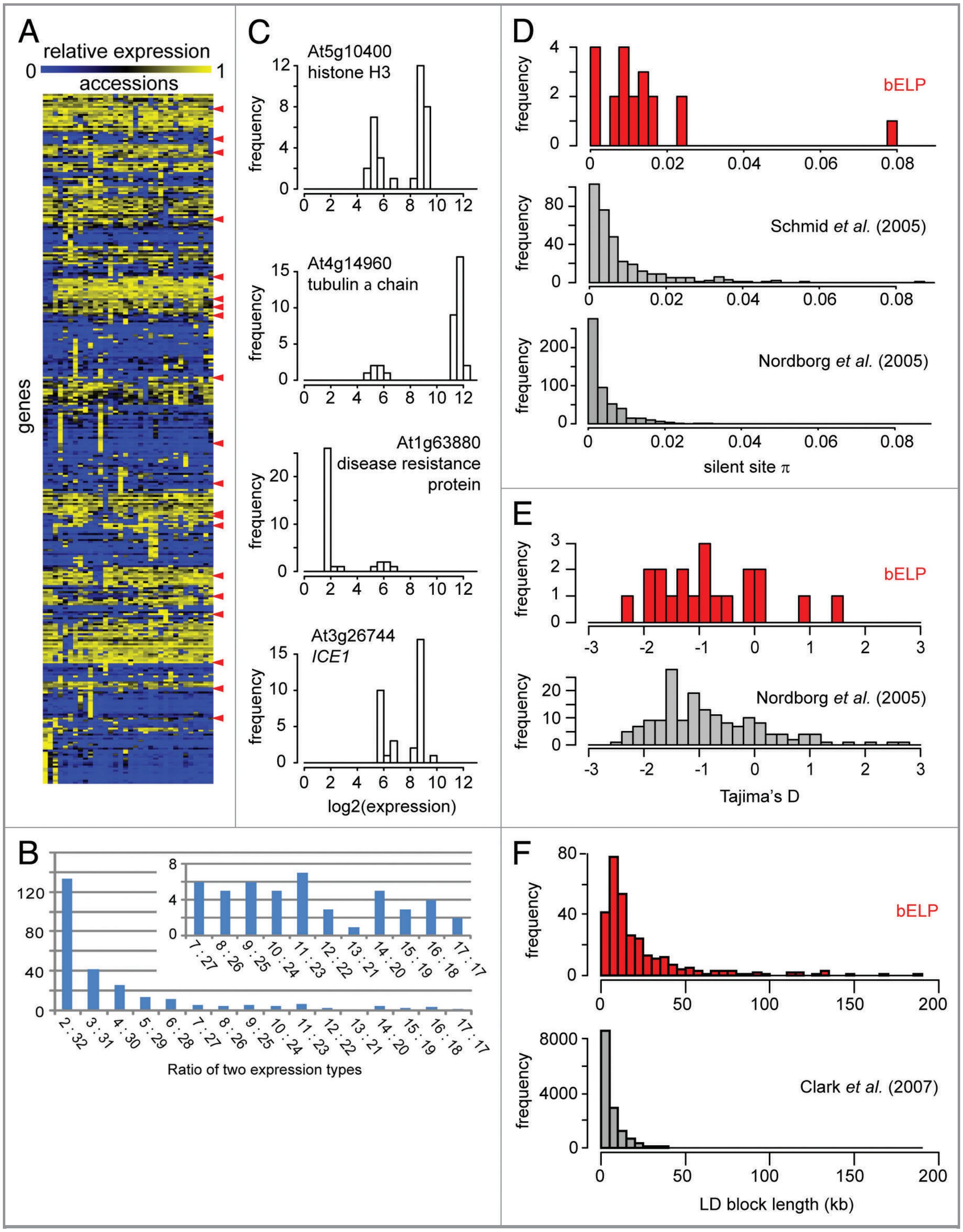

**Figure 2.** bELPs in *A. thaliana* and analyses of the selected sequences. (A) Heat map of bELP gene expression (dmax = 1). Low and high levels of expression are shown by blue and yellow colors, respectively. Thirty-four accessions are aligned horizontally and 356 genes are aligned vertically. Red arrowheads indicate the 20 sequenced genes. A full resolution image with AGI codes and annotations of bELP genes is available as **Figure S2**. (B) Frequency distribution of the ratio of two expression types. Inset, enlarged view. (C) Examples of bELP genes. Each histogram shows two expression modes. (D) Histogram of silent-site nucleotide diversity (π) of the 20 sequenced bELP genes (red, upper) compared with the genomic distribution (gray, two lower histograms). Data for the genomic distribution are from Schmid et al.[39] and Nordborg et al.[38] (E) Histogram of Tajima's D of the 20 sequenced bELP genes (red, upper) compared with the genomic distribution (gray, lower). Data for the genomic distribution are from Nordborg et al.[38] (F) Histograms of the lengths of LD blocks. Upper: LD blocks including bELP genes. Lower: genome-wide distribution estimated from Clark et al.[40] LD blocks over 250 kb (0.002%) are not shown in the lower histogram.

**Table 2.** Measures of diversity for the 20 bELP genes

| AGI code | $n^a$ | Length$^b$ (bp) | $S^c$ | $\pi^d$ | $\theta_w{}^d$ | Tajima's D | Fu and Li's D* |
|---|---|---|---|---|---|---|---|
| At1g11280 | 17 | 610 | 10 | 0.0024 | 0.0049 | -1.8776* | -2.6346* |
| At1g23780 | 23 | 577 | 114 | 0.0782 | 0.0569 | 1.5053 | 0.9226 |
| At1g52040 | 22 | 419 | 44 | 0.0157 | 0.0255 | -1.6706 | -2.3791 |
| At1g53690 | 23 | 544 | 24 | 0.0089 | 0.0121 | -0.9712 | 0.4331 |
| At1g58270 | 21 | 725 | 50 | 0.0227 | 0.0228 | -0.1620 | -0.0944 |
| At1g64190 | 21 | 549 | 36 | 0.0246 | 0.0194 | 0.9314 | 0.4395 |
| At1g73330 | 23 | 544 | 3 | 0.0005 | 0.0015 | -1.7313 | -2.4937 |
| At2g25450 | 8 | 432 | 13 | 0.0091 | 0.0127 | -1.4567 | -1.6851 |
| At2g40010 | 17 | 516 | 29 | 0.0083 | 0.0183 | -2.2182** | -3.0134* |
| At3g27200 | 20 | 474 | 31 | 0.0165 | 0.0187 | -0.5667 | -0.4396 |
| At3g26744 | 15 | 642 | 28 | 0.0102 | 0.0136 | -1.0486 | 0.0394 |
| At3g44280 | 12 | 592 | 13 | 0.0059 | 0.0077 | -0.9857 | -1.3601 |
| At4g03060 | 13 | 644 | 5 | 0.0013 | 0.0027 | -1.8631* | -2.3235* |
| At4g05050 | 24 | 700 | 40 | 0.0173 | 0.0170 | 0.0709 | -0.1278 |
| At4g22380 | 21 | 618 | 24 | 0.0115 | 0.0119 | -0.1277 | 0.0010 |
| At4g23810 | 15 | 539 | 6 | 0.0021 | 0.0035 | -1.3326 | -1.9560 |
| At5g01820 | 20 | 454 | 19 | 0.0128 | 0.0125 | 0.0983 | 0.4566 |
| At5g10400 | 18 | 498 | 13 | 0.0065 | 0.0084 | -0.8726 | -1.1086 |
| At5g44520 | 17 | 600 | 39 | 0.0133 | 0.0195 | -1.3134 | -2.0006 |
| At5g53940 | 22 | 603 | 21 | 0.0078 | 0.0097 | -0.7315 | -0.5220 |

**$p < 0.01$, *$p < 0.05$; [a]number of samples; [b]length of sequenced region; [c]number of segregation sites in the sample; [d]estimates are based on silent site.

accessions. Upstream regions and 5'UTRs of the genes were sequenced, because these regions may contain nucleotide polymorphisms responsible for bELPs. We obtained a total of 210,824 bp of nucleotide sequences and identified 526 non-redundant polymorphic sites (**Table 2**). Silent-site nucleotide diversities of the genes were highly variable ($\pi$ = 0.0005 ~0.0782; **Fig. 2D and Table 2**). Thirteen of the 20 sequenced bELP genes had greater silent-site nucleotide diversity than the mean level of 0.007 observed from previously studied *A. thaliana* genes.[37] A comparison of the distributions of the silent-site nucleotide diversity of bELP genes with that of the random genome-wide data set[38,39] suggested that bELP genes exhibited higher silent-site nucleotide diversity (**Fig. 2D**, $p < 0.01$, Mann-Whitney's U test). To assess whether high silent-site nucleotide diversity is a specific feature of bELP genes or a general feature of ELP genes, 344 genes exhibiting the most variable expression (but not exhibiting bELP) were collected and compared with bELP genes. As a result, we found that the bELP genes had significantly greater silent-site nucleotide diversity than did genes exhibiting expression polymorphisms but not bELPs ($p < 0.01$, Mann-Whitney's U test). However, with respect to Tajima's D statistics, there was no significant positive departure from a neutral-equilibrium model, although some genes exhibited significantly negative Tajima's D values (**Table 2**). The distribution of Tajima's D of bELP genes was not significantly different from that of the random genome-wide data set (**Fig. 2E**).

**Population genetic analyses based on previously published array-based resequence data.** In addition to the analyses of 20 bELP genes, we conducted further population genetic analyses based on the genome-wide polymorphism data of 20 accessions of *A. thaliana*.[40] First, we compared the lengths of the LD blocks including a bELP gene to the genome-wide distribution of the LD block lengths (**Fig. 2F**). The lengths of the LD blocks that included a bELP gene were highly variable (953 bp ~467,183 bp) and significantly longer than those of the genome-wide data (Mann-Whitney's U test; $p < 2.2 \times 10^{-16}$). We found that 92.2% of bELP genes were located on longer LD blocks than the median LD block length of genome-wide distribution (3,797 bp).

The $\pi$ of 4-fold degenerate sites (50-kb window) (provided by Clark et al.[40]) in regions that included bELP genes were compared with those distributed in genome-wide regions. Genomic regions, including bELPs, exhibited significantly greater $\pi$ values of 4-fold degenerate sites than the genome-wide data (**Fig. S4**; Mann-Whitney's U test; $p < 0.0045$).

**Sequencing analysis of At1g23780 and flanking regions.** One of the sequenced 20 bELP genes, At1g23780, exhibited an 11-fold greater silent-site nucleotide diversity ($\pi$ = 0.0782) than the genomic mean level. Although a high nucleotide diversity level is a signature of balancing selection, genetic hitchhiking may cause this diversity. To assess whether the At1g23780 gene was a target of the balancing selection, we performed further sequencing analysis that included the flanking genes (At1g23760 ~At1g23800). We obtained a total of 128,472 bp of nucleotide sequences of these five genes from 23 *A. thaliana* accessions and from three *Arabidopsis lyrata* individuals. For *A. thaliana* accessions, 292 non-redundant polymorphic sites were

**Table 3.** Measures of diversity for the genes flanking At1g23780

| AGI code | $n^a$ | Length$^b$ (bp) | S$^c$ | $\pi^d$ | $\theta_w{}^d$ | Tajima's D | Fu and Li's D* |
|---|---|---|---|---|---|---|---|
| At1g23760 | 23 | 590 | 5 | 0.0013 | 0.0041 | -1.81765* | -2.40051 |
| At1g23770 | 23 | 555 | 9 | 0.0059 | 0.0066 | -0.72468 | -0.7821 |
| At1g23780 | 23 | 2,820 | 238 | 0.0485 | 0.0341 | 1.71137 | 1.03643 |
| At1g23790 | 23 | 589 | 19 | 0.0218 | n.a. | 1.56202 | 0.70176 |
| At1g23800 | 23 | 585 | 21 | 0.0290 | 0.0168 | 2.57462** | 0.79733 |

**$p < 0.01$, *$p < 0.05$; $^a$number of samples; $^b$length of sequenced region; $^c$number of segregation sites in the sample; $^d$estimates are based on silent site.

determined (**Table 3**). Nucleotide diversities were calculated with a sliding window of 100 bp, in steps of 25 bp. The levels of nucleotide diversity from At1g23780 to At1g23800 were greater than the genomic mean level (**Fig. 3A**). A peak of nucleotide diversity was found at the At1g23780 locus and its upstream region, suggesting that the At1g23780 gene is a target of balancing selection. Neighbor-joining trees of the genes show that At1g23780, At1g23790 and At1g23800 fell into two major clades (**Fig. 3B**). The frequencies of the sequence groups were at intermediate levels. LD analyses based on the polymorphism data around At1g23780 and its flanking region (~35 kb), provided by Clark et al. (2007), showed that there is a high LD region between At1g23780 and At1g23800. These features indicate the presence of balancing selection involving the region containing At1g23780–800. Although At1g23790 and At1g23800 had two sequence groups, these genes had only one expression mode (**Fig. 3C**). Only At1g23780 exhibited bimodal expression. The expression modes and sequence groups of At1g23780 were correlated, except for two accessions (Van-0 and Kin-0 in **Fig. 4A**). The At1g23780 gene sequence of Van-0 is a chimera of Col-0-type and Cvi-type sequences. Single nucleotide polymorphisms (SNPs) among Col-0, Cvi and Van-0 are shown in **Figure 4B**. In the Van-0 accession, Col-0-type SNP alleles formed a cluster ($p < 1.4$ x 10–23). This result suggests that the sequence of Van-0 arose by two recombination events or by gene conversion. The promoter sequence of Van-0 was Cvi-type, although Van-0 exhibited a higher level of expression than Col-0. The recombinant region in Van-0 might affect its expression level. For Kin-0, we could not identify sequence differences that explained the differences in expression. A possible explanation is the existence of a trans-factor polymorphism. Polymorphism of a transcriptional factor controlling expression of At1g23780 might explain the expression level polymorphism in Kin-0. The product of the At1g23780 gene is annotated as a putative F-box protein. Because amino acid substitutions were not found in the F-box domain, the At1g23780 protein was expected to act as an F-box protein in every accession that we sequenced.

## Discussion

We found that the silent-site nucleotide diversity of bELP genes is significantly greater than that of the random genome-wide distribution (**Fig. 2D**). A greater nucleotide diversity is expected when the polymorphisms are maintained for long periods of time, which is considered to be a possible signature of balancing selection.[4] We also found that the lengths of the LD blocks that included bELP genes were significantly longer than those in the genome-wide distribution (**Fig. 2F**). A longer LD is expected when balancing selection occurs, e.g., a recent, partial selective sweep associated with local adaptation.[4] Our results suggest that balancing selection is occurring on the genomic regions containing bELP genes. bELPs could be useful for the screening of candidate genes targeted by balancing selection. The association between bELPs and genomic signatures of balancing selection suggests that the majority of bELPs are the result of cis-acting nucleotide polymorphisms (e.g., polymorphism of cis-acting regulatory region, splicing junction and whole gene deletion). This is consistent with previous reports of ELPs and of the molecular basis controlling ELPs.[15,41]

The bELP genes identified in this study include genes that were previously studied as targets of balancing selection. An example is ESP (**Table S3**), the product of which promotes herbivory by generalist herbivores and decreases oviposition by specialist herbivores. ESP is thought to be selected depending on the surrounding herbivore fauna.[27,42] Another example is the At1g63880 gene, which encodes a putative TIR-NBS-LRR-type disease resistance protein (**Fig. 2C**). At1g63880 is located in the high diversity region 1 defined in Cork et al., 2005, which was identified by analysis of a large-scale sequence data set. The genomic region around At1g63880 exhibited high nucleotide diversity ($\pi = 0.068$). The TIR-NBS-LRR gene cluster containing At1g63880 was suggested to be a target of balancing selection,[13] an example that suggests that ELP analysis helps to isolate selection targets from among candidate genes identified by genomic scans of nucleotide sequences.

The physiological roles of some bELP genes have been examined in previous studies. However, the modes of selection for such genes were not addressed. For example, inducer of CBF Expression 1 (ICE1) is a key transcriptional factor for controlling resistance to cold stress.[43-45] The expression of *ICE1* exhibited a clear bELP profile (**Fig. 2C**). This result raises the possibility that ICE1 has a role in phenotypic polymorphisms with respect to cold resistance. The ICE1 gene was also reported as the SCREAM (SCRM) gene, which determines successive initiation, proliferation, and terminal differentiation of stomatal cell lineages.[46] This surprising finding suggests a link between cold resistance and stomatal development. The pleiotropic function of ICE1 may play a role in maintaining its bELP status.

Our list of bELP genes includes some known to be responsible for phenotypic polymorphisms (i.e., FLC, ESP, AOP2, AOP3

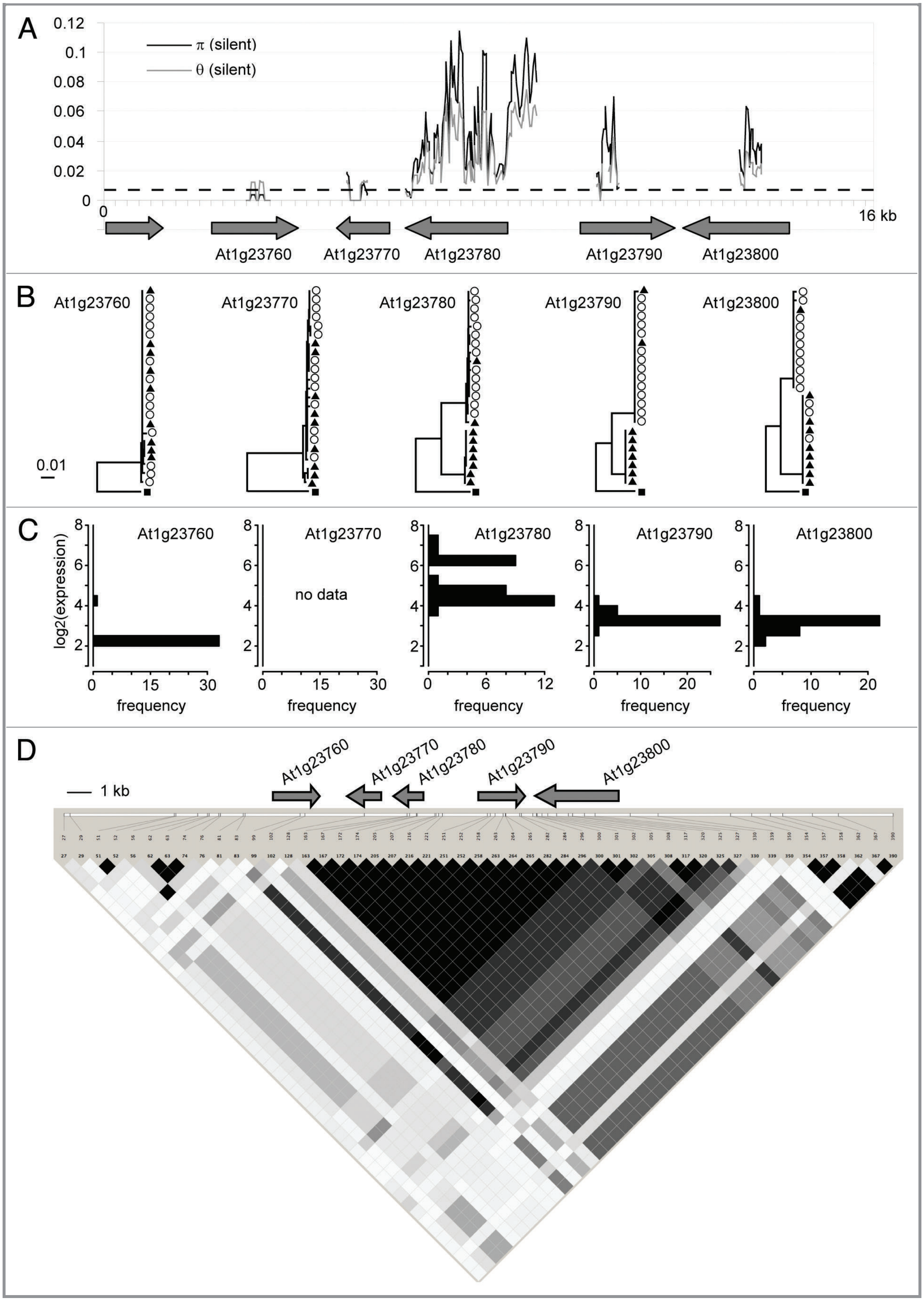
A
0.12
0.1
0.08
0.06
0.04
0.02
0
π (silent)
θ (silent)
0
16 kb
At1g23760
At1g23770
At1g23780
At1g23790
At1g23800
B
At1g23760
At1g23770
At1g23780
At1g23790
At1g23800
0.01
C
log2(expression)
8
6
4
2
At1g23760
0
15
30
frequency
At1g23770
no data
0
15
30
frequency
At1g23780
0
6
12
frequency
At1g23790
0
10
20
frequency
At1g23800
0
10
20
frequency
D
1 kb
At1g23760
At1g23770
At1g23780
At1g23790
At1g23800
27 29 51 52 56 62 63 74 76 81 83 99 102 128 163 167 172 174 205 207 216 221 251 252 258 263 264 265 282 284 296 300 301 302 305 308 317 320 325 327 330 339 350 354 357 358 362 367 390

**Figure 3 (See opposite page).** Nucleotide diversity, neighbor-joining trees and expression levels of At1g23780 and flanking genes. (A) Nucleotide diversity was calculated with a sliding window (100 bp). Gray arrows indicate genes. The dashed line indicates the average level of silent-site $\pi$ for *A. thaliana*. (B) Neighbor-joining trees of At1g23780 and flanking genes (from At1g23760 to At1g23800). Open circles and closed triangles indicate low and high expression accessions of the At1g23780 gene, respectively. Closed boxes represent *A. lyrata* as an outgroup. (C) Gene expression levels of At1g23780 and flanking genes. Data for At1g23770 are not available, because there was no corresponding probe-set on the ATH1 gene chip. (D) The $r^2$ plot of At1g23780 and its flanking region (~35 kb). Black indicates strong LD ($r^2$ = 1) and white indicates weak LD ($r^2$ = 0).

and RPP5). This result suggests that bELP analysis is effective in screening for genes associated with phenotypic polymorphisms. However, our list does not represent all of the genes responsible for phenotypic polymorphisms. One possible reason for the omission of some genes is that the phenotypic polymorphism is not derived from ELPs but from changes in protein function caused by amino acid substitutions. It is unclear to what degree balancing selection results from ELPs. A combination of large-scale resequence analysis[40] and systematic multi-conditional ELP analysis will reveal the quantitative importance of ELPs for balancing selection.

It is worth mentioning that the bELP genes include several genes responsible for fundamental cellular processes, e.g., histones, actin, tubulins, ubiquitin and ribosomal proteins (**Fig. 2C**; **Table S3**). These genes are expected to have a stable expression level among accessions, because they function in fundamental cellular processes and are ubiquitously expressed in various tissues and at different developmental stages. One possible reason is that they are present in multiple copies, which may compensate for the expression level of a specific polymorphic gene. Another possible reason is selection for unknown pleiotropic functions of the genes. These genes are commonly used as internal

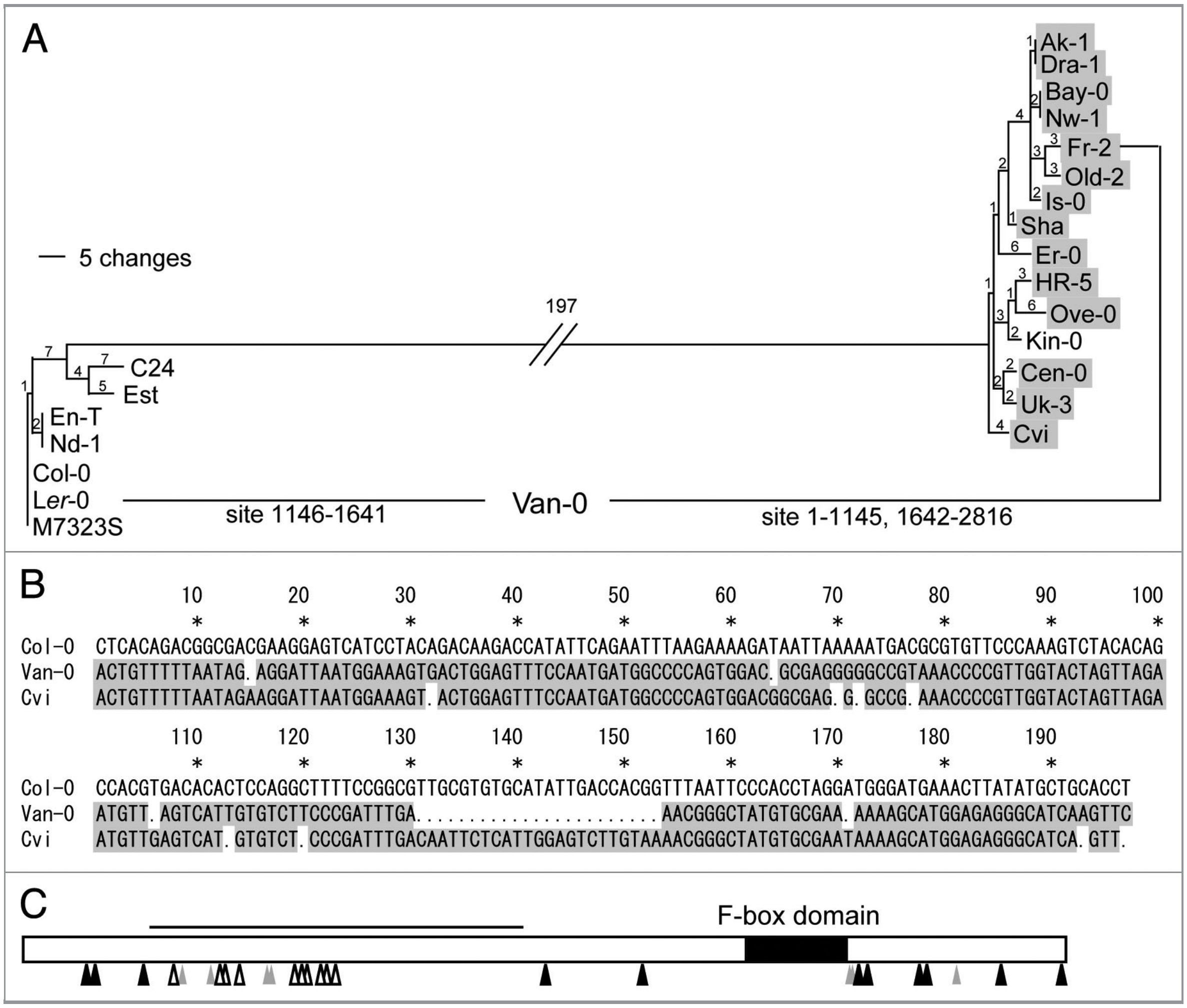
A
5 changes
197
Ak-1
Dra-1
Bay-0
Nw-1
Fr-2
Old-2
Is-0
Sha
Er-0
HR-5
Ove-0
Kin-0
Cen-0
Uk-3
Cvi
C24
Est
En-T
Nd-1
Col-0
Ler-0
M7323S
Van-0
site 1146-1641
site 1-1145, 1642-2816
B
10 20 30 40 50 60 70 80 90 100
Col-0 CTCACAGACGGCGACGAAGGAGTCATCCTACAGACAAGACCATATTCAGAATTTAAGAAAAGATAATTAAAAATGACGCGTGTTCCCAAAGTCTACACAG
Van-0 ACTGTTTTTAATAG.AGGATTAATGGAAAGTGACTGGAGTTTCCAATGATGGCCCCAGTGGAC.GCGAGGGGGCCGTAAACCCCGTTGGTACTAGTTAGA
Cvi ACTGTTTTTAATAGAAGGATTAATGGAAAGT.ACTGGAGTTTCCAATGATGGCCCCAGTGGACGGCGAG.G.GCCG.AAACCCCGTTGGTACTAGTTAGA
110 120 130 140 150 160 170 180 190
Col-0 CCACGTGACACACTCCAGGCTTTTCCGGCGTTGCGTGTGCATATTGACCACGGTTTAATTCCCACCTAGGATGGGATGAAACTTATATGCTGCACCT
Van-0 ATGTT.AGTCATTGTGTCTTCCCGATTTGA.......................AACGGGCTATGTGCGAA.AAAAGCATGGAGAGGGCATCAAGTTC
Cvi ATGTTGAGTCAT.GTGTCT.CCCGATTTGACAATTCTCATTGGAGTCTTGTAAAACGGGCTATGTGCGAATAAAAGCATGGAGAGGGCATCA.GTT.
C
F-box domain

**Figure 4.** Relationship between haplotypes of the At1g23780 gene. (A) The haplotype network for At1g23780. Names of accessions with low expression levels are shaded. Minimum steps are indicated for each node. (B) Col-0 and Cvi polymorphic nucleotides. Nucleotides also shared by Col-0 are depicted by dots. Nucleotides also shared by Cvi are shaded. (C) Schematic of the At1g23780 gene. The upper line indicates the putative recombined region in the Van-0 accession. Triangles indicate amino acid substitutions. (Closed triangles represent amino acid differences between the Col-0 and Cvi alleles; gray triangles indicate minor variations). The closed box indicates the F-box domain.

standards in expression analyses, because their expression is thought to be stable. This result suggests that care should be taken in selecting internal standard genes.

The At1g23780 gene was identified as a possible target of balancing selection. This gene encodes a putative F-box protein, a subunit of E3-type ubiquitin ligases, which are key components for defining which proteins are degraded by proteasomes. Protein degradation by proteasomes controls various biological processes, for example hormone reception, signal transduction and the defense response (LECHNER et al., 2006). It is not known in which biological processes the At1g23780 protein plays a role. Unfortunately, we could not obtain a knockout line of the At1g23780 gene. In addition to At1g23780, the gene family encoding F-box proteins exhibits high nucleotide diversity.[40] F-box proteins seem to be good candidates for the functional analysis of polymorphisms at the molecular level.

We found that bELP genes were generally located on genomic regions with significantly longer LDs than the genomic distribution. This result indicates that a flanking region of each bELP gene has a population genomic pattern similar to that of the respective bELP gene. Polymorphisms targeted by balancing selection in long LD regions are hard to identify by scans based on genomic sequences alone. In such cases, functional analysis of the balancing selection is needed to elucidate the molecular mechanisms underlying the selection. Which gene is involved in the selection? Which polymorphisms are involved in the selection? How do the polymorphisms affect the phenotype targeted by selection? We believe that ELPs provide good starting points for further functional analysis, because many ELPs are reported as causes of phenotypic variations[19,24,25,27] and because a measurement of expression level is usually easier to make than a biochemical and cell biological analysis of protein function. Indeed, we found a similar gene genealogy, a high nucleotide diversity and a linkage block from At1g23780 to At1g23800 (**Fig. 3A, B and D**). Only At1g23780 showed a clear bELP pattern in this region (**Fig. 3C**). While identifying ELP is not conclusive evidence that the gene is the target of selection, it will be worth conducting a further functional analysis of At1g23780. Such "multi-omics" approaches have the potential to facilitate the scanning of genes relevant for balanced polymorphisms not only in A. thaliana but also in other model and non-model organisms.

## Materials and Methods

**Screening for bimodal expression level polymorphisms.** Expression data were obtained from Detlef Weigel's website (www.weigelworld.org/resources/microarray/AtGenExpress/). The data were generated as a part of the AtGenExpress project and were also deposited in TAIR as ME00374 AtGenExpress: Ecotypes (triplicated data from 10 accessions) and ME00375 AtGenExpress: Ecotypes singletons (non-replicated data from 24 accessions).[22] Triplicate ME00374 data were averaged and used for fitting a mixture distribution of two normal distributions and screening by the simple nonparametric approach. Data were normalized to the average expression levels of each accession. ANOVA was conducted using normalized data of ME00374. The correction of multiple testing by ANOVA was conducted by controlling the false discovery rate (FDR = 0.01).[47] The estimation of a mixture distribution of two normal distributions from expression data was conducted using the vglm (family = mix2normal) function in the VGAM package of R.[48,49] In the simple nonparametric approach, bELPs were screened with bimodal3.pl, a Perl script developed in-house. Procedures to assess whether a given gene was a bELP gene are described below. First, the expression values of each accession were ordered and a progression of differences was calculated. Next, the maximum value for the progression was designated as $d_{max}$ (**Fig. 1B**). Finally, $d_{max}$ was compared with a threshold. If the $d_{max}$ was larger than the threshold, the gene was treated as a bELP gene. Based on the screening method, the term "bELP" was defined in this study as a gene that exhibits at least two modes of expression. The Perl script for screening is available upon request. The statistical analysis of the gene ontology was performed with GeneSpring GX (Agilent Technologies).

**Plant materials and growth conditions.** *A. thaliana* accessions (**Table S1**) were from Europe, North America, Central Asia and North Africa. *A. thaliana* seeds were obtained from ABRC. *Arabidopsis lyrata* (W313) seeds were provided by Kentaro K. Shimizu (University of Zurich). Seeds were surface-sterilized and then sown onto 0.5% (w/v) Gellan Gum (Wako) with MS medium (Wako) and 1% (w/v) sucrose and grown at 22°C under continuous light conditions.

**Sequencing.** Genomic DNA was isolated from the aerial parts of three *A. lyrata* plants and 23 *A. thaliana* accessions. PCR primers were designed from Col-0 genomic sequences (AGI 2000) using Primer3 (ROZEN and SKALETSKY 2000). Primer sequences are described in **Table S2**. Some *A. lyrata* primers were designed from the *A. lyrata* sequences generated in this study (**Table S2**). PCR was performed with ExTaq DNA polymerase (Takara), and amplified PCR products were purified using MagExtractor (TOYOBO). *A. thaliana* sequences were determined by direct sequencing of amplified PCR products. Purified PCR products of the *A. lyrata* sample were cloned into the T7Blue vector (Merck, Darmstadt) using the DNA Ligation Kit ver.2 (Takara). Plasmids were isolated using QIAprep (Qiagen). At least four independent clones were sequenced. Sequence data from this article have been deposited into the DDBJ/EMBL/GenBank Data Libraries under accession nos. AB498072-AB498554.

**Molecular population genetic analysis and phylogenetic analysis.** Sequences were initially aligned using Clustal W.[50] The alignment was manually checked and corrected. Calculations of population genetic parameters were performed with DnaSP 4.10.9.[51] Silent-site nucleotide diversity was estimated as both $\pi$[52] and $\theta_w$.[53] Tajima's D[54] and Fu and Li's D*[55] were also estimated. Phylogenetic analysis was performed using PAUP4.0*b10[56] to decipher the relationship between the ELPs of the highly polymorphic gene At1g23780 and its nucleotide sequence. The empirical genomic distributions of $\pi$ and Tajima's D in the silent site and intergenic regions were estimated from published sequence data.[38] We estimated $\pi$ and Tajima's D of populations consisting of *A. thaliana* accessions that were used in this article for bELP screening and that were sequenced in Nordborg et al.[38] While almost all of the genes described in Nordborg et al.[38] were used for estimation of $\pi$ and Tajima's D, the following nucleotide sites were excluded: those with alignment gaps and missing data, and all sites in gene 22387873 on chromosome 1, because the number of nucleotides in the sequence file and in the annotation file were mismatched.

**Linkage disequilibrium and recombination analysis.** For LD analyses, polymorphism data provided by the Perlegen oligonucleotide array of 20 accessions was used.[40] While Clark et al.[40] proposed several algorithms for SNP detection, SNPs called "MBML2" were used in our analyses. The following sites were excluded from the analyses: SNPs containing at least one accession that was not detected by the "MBML2" and SNPs that were singletons. The length of an LD block was defined as the distance between the first flanking SNP on the downstream side of the focal polymorphism that fails the four-gamete test[57] and the focal polymorphism. Note that block length based on the four-gamete test creates a bias for a relatively longer estimate, because the four-gamete test is an estimate of the minimum number of recombinations. The r2 plot of At1g23780 and its flanking region was generated by Haploview 4.0.[58] Several bELP genes were excluded from the analyses. These were unannotated genes and genes located at the beginning or the end of a chromosome, where there were too few markers to estimate the lengths of the LD blocks. The histogram presented in **Figure 2F** was developed not on a gene-wise basis, but on a block-wise basis, because sometimes more than one bELP gene was found on the same LD block. Finally, 306 out of 344 bELP genes (**Fig. 2F**) were used for making comparisons of distribution by the Mann-Whitney's U-test.

### Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

### Acknowledgments

We thank Sigeo Sugano for his assistance in sequence alignment, Shigeto Dobata for his advice about statistics and Kentaro K. Shimizu, Tetsuhiro Kawagoe and Shin-ichi Morinaga for their critical comments on the manuscript. PCR, purification and sequencing were performed in the Biotechnology Center of the University of Akita Prefecture. This work was supported by Core Research for Evolutional Science and Technology of the Japan Science and Technology Corporation and by Grants-in-Aid for Scientific Research (nos. 16044224, 16085203, 17107002) and for 21st Century Centers of Excellence Research, Kyoto University (A14) from the Ministry of Education, Culture, Sports, Science and Technology of Japan and for research fellowships to A.N., Y.O. and T.T. from the Japan Society for the Promotion of Science.

### Supplemental Material

Supplemental materials may be found here:
www.landesbioscience.com/journals/psb/article/20486

### References

1. Charlesworth B, Miyo T, Borthwick H. Selection responses of means and inbreeding depression for female fecundity in Drosophila melanogaster suggest contributions from intermediate-frequency alleles to quantitative trait variation. Genet Res 2007; 89:85-91; PMID:17521472; http://dx.doi.org/10.1017/S001667230700866X
2. Hartl DL, Clark AG. Principles of population genetics. 2007; Sinauuer Associates, Inc.
3. Hedrick P. Genetic polymorphism in heterogeneous environments: The age of genomics. Annu Rev Ecol Evol Syst 2006; 37:67-93; http://dx.doi.org/10.1146/annurev.ecolsys.37.091305.110132
4. Charlesworth D. Balancing selection and its effects on sequences in nearby genome regions. PLoS Genet 2006; 2:e64; PMID:16683038; http://dx.doi.org/10.1371/journal.pgen.0020064
5. Garrigan D, Hedrick PW. Perspective: detecting adaptive molecular polymorphism: lessons from the MHC. Evolution 2003; 57:1707-22; PMID:14503614; http://dx.doi.org/10.1111/j.0014-3820.2003.tb00580.x
6. Kreitman ME, Aguadé M. Excess polymorphism at the Adh locus in Drosophila melanogaster. Genetics 1986; 114:93-110; PMID:3021568
7. Kreitman M, Hudson RR. Inferring the evolutionary histories of the Adh and Adh-dup loci in Drosophila melanogaster from patterns of polymorphism and divergence. Genetics 1991; 127:565-82; PMID:1673107
8. Kroymann J, Donnerhacke S, Schnabelrauch D, Mitchell-Olds T. Evolutionary dynamics of an Arabidopsis insect resistance quantitative trait locus. Proc Natl Acad Sci U S A 2003; 100(Suppl 2):14587-92; PMID:14506289; http://dx.doi.org/10.1073/pnas.1734046100
9. Bakker EG, Toomajian C, Kreitman M, Bergelson J. A genome-wide survey of R gene polymorphisms in Arabidopsis. Plant Cell 2006; 18:1803-18; PMID:16798885; http://dx.doi.org/10.1105/tpc.106.042614
10. Bergelson J, Kreitman M, Stahl EA, Tian D. Evolutionary dynamics of plant R-genes. Science 2001; 292:2281-5; PMID:11423651; http://dx.doi.org/10.1126/science.1061337
11. Stahl EA, Dwyer G, Mauricio R, Kreitman M, Bergelson J. Dynamics of disease resistance polymorphism at the Rpm1 locus of Arabidopsis. Nature 1999; 400:667-71; PMID:10458161; http://dx.doi.org/10.1038/23260
12. Tian D, Araki H, Stahl E, Bergelson J, Kreitman M. Signature of balancing selection in Arabidopsis. Proc Natl Acad Sci U S A 2002; 99:11525-30; PMID:12172007; http://dx.doi.org/10.1073/pnas.172203599
13. Cork JM, Purugganan MD. High-diversity genes in the Arabidopsis genome. Genetics 2005; 170:1897-911; PMID:15911589; http://dx.doi.org/10.1534/genetics.104.036152
14. Reininga JM, Nielsen D, Purugganan MD. Functional and geographical differentiation of candidate balanced polymorphisms in Arabidopsis thaliana. Mol Ecol 2009; 18:2844-55; PMID:19457201; http://dx.doi.org/10.1111/j.1365-294X.2009.04206.x
15. Kliebenstein DJ, West MA, van Leeuwen H, Kim K, Doerge RW, Michelmore RW, et al. Genomic survey of gene expression diversity in Arabidopsis thaliana. Genetics 2006; 172:1179-89; PMID:16204207; http://dx.doi.org/10.1534/genetics.105.049353
16. Doerge RW. Mapping and analysis of quantitative trait loci in experimental populations. Nat Rev Genet 2002; 3:43-52; PMID:11823790; http://dx.doi.org/10.1038/nrg703
17. Carroll SB. Endless forms: the evolution of gene regulation and morphological diversity. Cell 2000; 101:577-80; PMID:10892643; http://dx.doi.org/10.1016/S0092-8674(00)80868-5
18. Wang H, Nussbaum-Wagler T, Li B, Zhao Q, Vigouroux Y, Faller M, et al. The origin of the naked grains of maize. Nature 2005; 436:714-9; PMID:16079849; http://dx.doi.org/10.1038/nature03863
19. Konishi S, Izawa T, Lin SY, Ebana K, Fukuta Y, Sasaki T, et al. An SNP caused loss of seed shattering during rice domestication. Science 2006; 312:1392-6; PMID:16614172; http://dx.doi.org/10.1126/science.1126410
20. Caicedo AL, Stinchcombe JR, Olsen KM, Schmitt J, Purugganan MD. Epistatic interaction between Arabidopsis FRI and FLC flowering time genes generates a latitudinal cline in a life history trait. Proc Natl Acad Sci U S A 2004; 101:15670-5; PMID:15505218; http://dx.doi.org/10.1073/pnas.0406232101
21. Johanson U, West J, Lister C, Michaels S, Amasino R, Dean C. Molecular analysis of FRIGIDA, a major determinant of natural variation in Arabidopsis flowering time. Science 2000; 290:344-7; PMID:11030654; http://dx.doi.org/10.1126/science.290.5490.344
22. Lempe J, Balasubramanian S, Sureshkumar S, Singh A, Schmid M, Weigel D. Diversity of flowering responses in wild Arabidopsis thaliana strains. PLoS Genet 2005; 1:109-18; PMID:16103920; http://dx.doi.org/10.1371/journal.pgen.0010006
23. Gassmann W, Hinsch ME, Staskawicz BJ. The Arabidopsis RPS4 bacterial-resistance gene is a member of the TIR-NBS-LRR family of disease-resistance genes. Plant J 1999; 20:265-77; PMID:10571887; http://dx.doi.org/10.1046/j.1365-313X.1999.t01-1-00600.x

24. Grant MR, Godiard L, Straube E, Ashfield T, Lewald J, Sattler A, et al. Structure of the Arabidopsis RPM1 gene enabling dual specificity disease resistance. Science 1995; 269:843-6; PMID:7638602; http://dx.doi.org/10.1126/science.7638602
25. Kliebenstein DJ, Lambrix VM, Reichelt M, Gershenzon J, Mitchell-Olds T. Gene duplication in the diversification of secondary metabolism: tandem 2-oxoglutarate-dependent dioxygenases control glucosinolate biosynthesis in Arabidopsis. Plant Cell 2001; 13:681-93; PMID:11251105; http://dx.doi.org/10.1105/tpc.13.3.681
26. Kliebenstein D, Pedersen D, Barker B, Mitchell-Olds T. Comparative analysis of quantitative trait loci controlling glucosinolates, myrosinase and insect resistance in Arabidopsis thaliana. Genetics 2002; 161:325-32; PMID:12019246
27. Lambrix V, Reichelt M, Mitchell-Olds T, Kliebenstein DJ, Gershenzon J. The Arabidopsis epithiospecifier protein promotes the hydrolysis of glucosinolates to nitriles and influences Trichoplusia ni herbivory. Plant Cell 2001; 13:2793-807; PMID:11752388; http://dx.doi.org/10.1105/tpc.010261
28. Ellegren H, Sheldon BC. Genetic basis of fitness differences in natural populations. Nature 2008; 452:169-75; PMID:18337813; http://dx.doi.org/10.1038/nature06737
29. Mitchell-Olds T, Schmitt J. Genetic mechanisms and evolutionary significance of natural variation in Arabidopsis. Nature 2006; 441:947-52; PMID:16791187; http://dx.doi.org/10.1038/nature04878
30. Stinchcombe JR, Hoekstra HE. Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. Heredity (Edinb) 2008; 100:158-70; PMID:17314923; http://dx.doi.org/10.1038/sj.hdy.6800937
31. Weigel D, Nordborg M. Natural variation in Arabidopsis. How do we find the causal genes? Plant Physiol 2005; 138:567-8; PMID:15955918; http://dx.doi.org/10.1104/pp.104.900157
32. Teng S, Keurentjes J, Bentsink L, Koornneef M, Smeekens S. Sucrose-specific induction of anthocyanin biosynthesis in Arabidopsis requires the MYB75/PAP1 gene. Plant Physiol 2005; 139:1840-52; PMID:16299184; http://dx.doi.org/10.1104/pp.105.066688
33. Sønderby IE, Hansen BG, Bjarnholt N, Ticconi C, Halkier BA, Kliebenstein DJ. A systems biology approach identifies a R2R3 MYB gene subfamily with distinct and overlapping functions in regulation of aliphatic glucosinolates. PLoS One 2007; 2:e1322; PMID:18094747; http://dx.doi.org/10.1371/journal.pone.0001322
34. Brem RB, Yvert G, Clinton R, Kruglyak L. Genetic dissection of transcriptional regulation in budding yeast. Science 2002; 296:752-5; PMID:11923494; http://dx.doi.org/10.1126/science.1069516
35. Keurentjes JJ, Fu J, Terpstra IR, Garcia JM, van den Ackerveken G, Snoek LB, et al. Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. Proc Natl Acad Sci U S A 2007; 104:1708-13; PMID:17237218; http://dx.doi.org/10.1073/pnas.0610429104
36. West MA, Kim K, Kliebenstein DJ, van Leeuwen H, Michelmore RW, Doerge RW, et al. Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in Arabidopsis. Genetics 2007; 175:1441-50; PMID:17179097; http://dx.doi.org/10.1534/genetics.106.064972
37. Yoshida K, Kamiya T, Kawabe A, Miyashita NT. DNA polymorphism at the ACAULIS5 locus of the wild plant Arabidopsis thaliana. Genes Genet Syst 2003; 78:11-21; PMID:12655134; http://dx.doi.org/10.1266/ggs.78.11
38. Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, Zheng H, et al. The pattern of polymorphism in Arabidopsis thaliana. PLoS Biol 2005; 3:e196; PMID:15907155; http://dx.doi.org/10.1371/journal.pbio.0030196
39. Schmid KJ, Ramos-Onsins S, Ringys-Beckstein H, Weisshaar B, Mitchell-Olds T. A multilocus sequence survey in Arabidopsis thaliana reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. Genetics 2005; 169:1601-15; PMID:15654111; http://dx.doi.org/10.1534/genetics.104.033795
40. Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G, Shinn P, et al. Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. Science 2007; 317:338-42; PMID:17641193; http://dx.doi.org/10.1126/science.1138632
41. Plantegenet S, Weber J, Goldstein DR, Zeller G, Nussbaumer C, Thomas J, et al. Comprehensive analysis of Arabidopsis expression level polymorphisms with simple inheritance. Mol Syst Biol 2009; 5:242; PMID:19225455; http://dx.doi.org/10.1038/msb.2008.79
42. Kliebenstein DJ, Kroymann J, Mitchell-Olds T. The glucosinolate-myrosinase system in an ecological and evolutionary context. Curr Opin Plant Biol 2005; 8:264-71; PMID:15860423; http://dx.doi.org/10.1016/j.pbi.2005.03.002
43. Chinnusamy V, Ohta M, Kanrar S, Lee BH, Hong X, Agarwal M, et al. ICE1: a regulator of cold-induced transcriptome and freezing tolerance in Arabidopsis. Genes Dev 2003; 17:1043-54; PMID:12672693; http://dx.doi.org/10.1101/gad.1077503
44. Lee BH, Henderson DA, Zhu JK. The Arabidopsis cold-responsive transcriptome and its regulation by ICE1. Plant Cell 2005; 17:3155-75; PMID:16214899; http://dx.doi.org/10.1105/tpc.105.035568
45. Miura K, Jin JB, Lee J, Yoo CY, Stirm V, Miura T, et al. SIZ1-mediated sumoylation of ICE1 controls CBF3/DREB1A expression and freezing tolerance in Arabidopsis. Plant Cell 2007; 19:1403-14; PMID:17416732; http://dx.doi.org/10.1105/tpc.106.048397
46. Kanaoka MM, Pillitteri LJ, Fujii H, Yoshida Y, Bogenschutz NL, Takabayashi J, et al. SCREAM/ICE1 and SCREAM2 specify three cell-state transitional steps leading to arabidopsis stomatal differentiation. Plant Cell 2008; 20:1775-85; PMID:18641265; http://dx.doi.org/10.1105/tpc.108.060848
47. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc, B 1995; 57:289-300; http://dx.doi.org/10.2307/2346101
48. R. D. C. T. R. A language and environment for statistical computing. R Foundation for Statistical Computing 2011.
49. Yee T. The VGAM Package for Categorical Data Analysis. J Stat Softw 2010; 32:1-34.
50. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994; 22:4673-80; PMID:7984417; http://dx.doi.org/10.1093/nar/22.22.4673
51. Rozas J, Sánchez-DelBarrio JC, Messeguer X, Rozas R, Dna SP. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 2003; 19:2496-7; PMID:14668244; http://dx.doi.org/10.1093/bioinformatics/btg359
52. Tajima F. Evolutionary relationship of DNA sequences in finite populations. Genetics 1983; 105:437-60; PMID:6628982
53. Watterson GA. On the number of segregating sites in genetical models without recombination. Theor Popul Biol 1975; 7:256-76; PMID:1145509; http://dx.doi.org/10.1016/0040-5809(75)90020-9
54. Tajima F. Statistical analysis of DNA polymorphism. Jpn J Genet 1993; 68:567-95; PMID:8031577; http://dx.doi.org/10.1266/jjg.68.567
55. Fu YX, Li WH. Statistical tests of neutrality of mutations. Genetics 1993; 133:693-709; PMID:8454210
56. Swofford DL. PAUP*: phylogenetic analysis using parsimony (*and other methods), version 4. 2002; Sinauuer Associates, Inc.
57. Hudson RR, Kaplan NL. Deleterious background selection with recombination. Genetics 1995; 141:1605-17; PMID:8601498
58. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 2005; 21:263-5; PMID:15297300; http://dx.doi.org/10.1093/bioinformatics/bth457